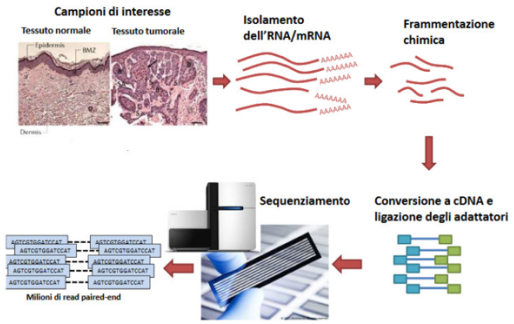
Rna-Seq e analisi bioinformatica

Il **trascrittoma** è l'insieme delle molecole di mRNA (o trascritti) presenti in una cellula. Un'applicazione della trascrittomica quantitativa è l'analisi differenziale dell'espressione genica, ottenuta confrontando i profili trascrizionali di due o più individui, tessuti o tipi cellulari. L'analisi del trascrittoma può avvenire secondo due tecnologie: quelle basate su ibridazione e quelle, più recenti, basate su sequenziamento.

Ibridazione	***Sequenziamento***
L'ibridazione si basa sulla proprietà dei nucleotidi di appaiarsi con i loro complementari fissati su un supporto. *Di questa categoria fanno parte i microarrays, da anni largamente utilizzati per ottenere informazioni sull'espressione genica.* *Sono costituiti da un supporto solido a cui sono ancorate delle sonde di DNA, dette probe, in numero molto elevato per ogni gene e disposte in posizioni note.* *L'RNA estratto dalla cellula viene retrotrascritto, marcato con una particella fluorescente e ibridato con il microarray. L'intensità della fluorescenza è una misura di quante molecole hanno ibridato il probe, ovvero di quanto il gene associato al probe è espresso nella cellula.* *Questa tecnica presenta numerosi limiti: la necessità di conoscere a priori le sequenze geniche per la progettazione dei probe e il limitato range dinamico (cioè il rapporto fra i livelli di massima e minima espressione genica misurabili) dovuto al rumore di fondo e al fenomeno di saturazione del segnale.*	*Per sequenziamento si intende, invece, l'identificazione della sequenza di DNA fornita in input alla strumentazione. Una tecnica recente per la misura del trascrittoma basata sul sequenziamento è l'RNA Sequencing (RNA- Seq), che si fonda sulle tecnologie di sequenziamento NGS (Next Generation Sequencing). Il protocollo di un esperimento di RNA-Seq varia in base alla tecnologia utilizzata, ma è comunque possibile descriverne in linea generale i passaggi principali: I campioni di mRNA sono estratti dalle cellule in analisi e preparati al sequenziamento separatamente: ciascuno campione viene frammentato casualmente in sequenze di dimensione inferiore, compre- sa tra i 200 e i 500 bp, per ottenere frammenti di dimensione compatibile con i sequenziatori in uso. Il processo di frammentazione è realizzato tramite idrolisi o nebulizzazione.* *Ciascun frammento viene poi retrotrascritto in DNA (cDNA). La retrotra-scrizione (o trascrizione inversa) è il processo di sintesi di un filamento di DNA complementare a partire da un filamento di RNA. Questo procedimento è eseguito al fine di aumentare la stabilità della molecola. Successivamente ad ogni frammento di cDNA vengono legate delle sequenze specifiche, chiamate*

	adattatori, che dipendono dal tipo di sequenziatore e che servono in varie fasi della preparazione del campione. *Le molecole di cDNA vengono amplificate mediante Polymerase Chain Reaction (PCR), procedimento che moltiplica il numero di copie di ciascun frammento per aumentarne la massa critica, e infine vengono fornite in input ai sequenziatori.* *Il DNA ottenuto viene sequenziato ad alto throughput per ottenere frammenti a sequenza nota, detti reads, di lunghezza variabile in base alla tecno- logia del sequenziatore utilizzato. Le reads sono quindi le sequenze, ottenute dal sequenziatore, che identificano l'ordine in cui si susseguono le basi nei frammenti di DNA. Esistono diversi tipi di sequenziatori che differiscono per le tecniche di sequenziamento, il tempo impiegato per le analisi, la lunghezza delle reads prodotte, la quantità di reads prodotte e la percentuale di errore per ogni run.* 

la tecnologia RNA-Seq offre diversi vantaggi che includono un range di livelli di espressione più ampio, un elevato throughput, un miglior coverage del genoma, rumore di fondo inferiore e la possibilità di esplorare nuovi trascritti di cui non sono note le sequenze. Per queste ragioni, l'RNA-Seq è pronto a rimpiazzare la tecnologia microarray e a diventare il principale strumento per la quantificazione dell'e- spressione genica, favorito inoltre dalla diminuzione dei tempi e dei costi di sequenziamento derivanti dallo sviluppo di sequenziatori di nuova generazio- ne (Next Generation Sequencing).

Tecnologia RNA-Seq:

1. **Esperimento RNA-Seq**:
    - Le reads rappresentano i dati grezzi per valutare l'espressione genica.
    - Maggiore è il numero di copie di un trascritto, maggiore è la probabilità di sequenziamento.
2. **Mappatura delle reads**:
    - Le reads sono allineate su un genoma o un trascrittoma di riferimento.
    - La scelta del riferimento (genoma o trascrittoma) influisce sugli algoritmi di allineamento.
3. **Genoma vs Trascrittoma**:
    - **Genoma**:
        - Composto da DNA intra-genico (67.5%) e DNA genico (37.5%).
        - Contiene introni (non codificanti) ed esoni (codificanti).
        - L'algoritmo di allineamento deve gestire grandi spazi (introni) tra le basi delle reads (Spliced Aligners).
    - **Trascrittoma**:
        - Costituito da mRNA maturi senza introni.
        - Algoritmi di allineamento non devono gestire lunghi spazi tra le basi delle reads (Unspliced Aligners).
4. **Criticità dell'allineamento**:
    - Identificazione della posizione migliore nel genoma o trascrittoma nonostante errori di sequenziamento e differenze tra campione e riferimento.
5. **Conteggio delle reads**:
    - Le reads allineate su un gene o trascritto vengono contate.
    - **Problema delle isoforme**:
        - Un gene può produrre diversi trascritti (isoforme) tramite splicing alternativo.
        - Ambiguità nel conteggio delle reads che mappano su esoni comuni a diverse isoforme.
6. **Conte (Counts)**:
    - Misura del livello di espressione di una regione del genoma.
    - Nei confronti comparativi, le espressioni di ciascun campione sono riassunte in una matrice.
    - **Matrice K**:
        - Dimensione G x m, dove G è il numero di regioni di interesse e m il numero di esperimenti.
        - Elemento Kij rappresenta il numero di reads mappate sulla variabile i nell'esperimento j.
7. **Modelli statistici per i counts**:
    - I counts sono descritti da variabili aleatorie seguendo modelli di distribuzione.
    - **Modelli più utilizzati**:
        - Modello di Poisson
        - Modello Binomiale Negativo

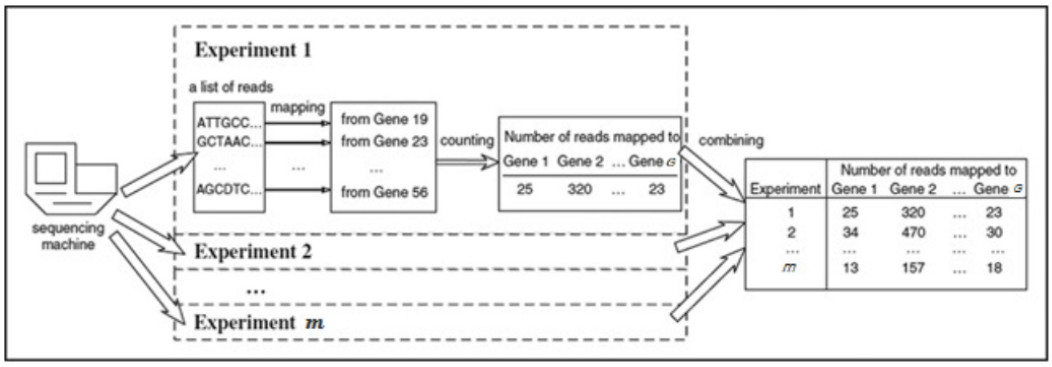


Figura 1.4: Procedimento per usare dati di RNA-Seq per analisi comparative.

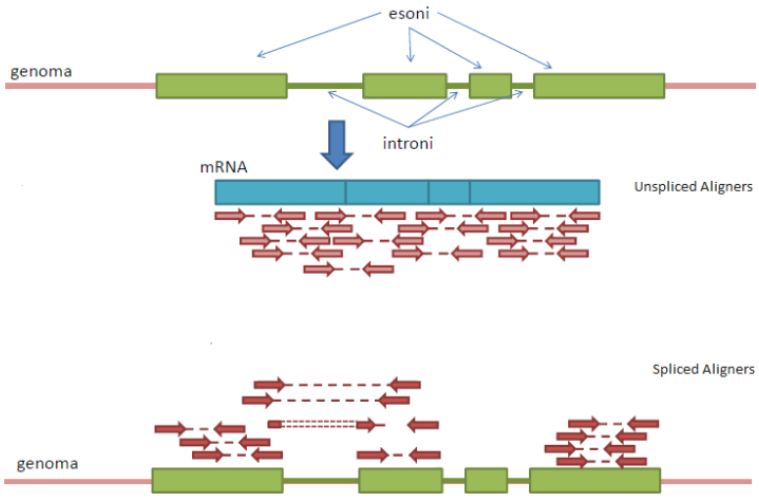


Figura 1.3: Esempio di allineamento al trascrtittoma e al genoma. Ciascuna coppia di frecce rosse indica una read.

ANALISI

Una volta ottenuta la matrice delle conte grezze, si passa alla fase di ana- lisi della differenziale espressione. Prima però, è necessario applicare delle trasformazioni ai dati al fine di risolvere alcune problematiche che sono in- trinseche del tipo di dato con cui si sta lavorando e del modo in cui questo è stato ottenuto. I passi per l'analisi di una matrice di conte di RNA-Seq pos- sono variare a seconda delle esigenze dei dati, ma in linea generale i passaggi principali da seguire sono:

1. Filtraggio;
2. Normalizzazione;
3. Identificazione dei geni differenzialmente espressi

Motivazione del filtraggio:

- **Obiettivo Teorico**: Tenere tutti i geni nell'analisi.
- **Pratica**: Filtrare i geni scarsamente espressi per migliorare l'affidabilità delle stime di espressione.

Problema dei geni scarsamente espressi:

- **Affidabilità**: La stima dell'espressione è meno affidabile per geni con conte basse.
- **Disturbo**: Questi geni possono influenzare negativamente la sensibilità e specificità dei metodi di analisi di differenziale espressione.

Scopo del filtraggio:

- **Rimuovere geni con bassa intensità o variabilità**: Questi geni sono meno probabili a fornire informazioni utili sul fenotipo d'interesse.

Criteri di filtraggio:

- **Esempio**: Eliminare geni con somma totale delle conte su tutti i campioni inferiore a 10.

Riduzione della perdita di potenza:

- **Test Multipli**: La perdita di potenza dovuta all'aggiustamento per test multipli può essere ridotta omettendo geni con scarse possibilità di essere rilevati come DE.
- **Condizione**: Il criterio di omissione deve essere indipendente dalla statistica test sotto l'ipotesi nulla.

Ottimizzazione del numero di test:

- **Obiettivo**: Mantenere il numero di test basso ma includere i geni d'interesse.
- **Risultato**: Se i geni differenzialmente espressi sono sovra-rappresentati nel filtraggio, il false discovery rate (FDR) associato a una certa soglia del test statistico sarà più basso.

Scopo della normalizzazione

- Rimuovere effetti sistematici dovuti alla tecnologia, assicurando che gli artefatti tecnici abbiano un minimo impatto sui risultati.

Fonti di variazione sistematica

1. **Variazione nella composizione dei nucleotidi**:
    - Il coverage delle reads potrebbe non essere uniforme lungo il genoma.
    - Geni lunghi avranno più reads associate rispetto a geni corti a parità di espressione.

2. **Bias tra campioni ("between-sample" biases)**:
    - **Profondità di sequenziamento**: Differente numero totale di reads mappate tra campioni, rendendo le conte osservate non direttamente comparabili.
    - **Riscalare le conte**: Risolvere la diversa dimensione delle librerie rendendo le dimensioni delle librerie equivalenti tra i campioni.

Problemi di riscalamento semplice

- **Geni altamente espressi**: Pochi geni altamente espressi possono catturare la maggior parte delle reads, causando una distribuzione ineguale delle reads tra gli altri geni.

Schemi di normalizzazione complessi

1. **Global Normalization**:
    - **Total Count (TC)**
    - **Upper Quartile (UQ)**
    - **Median (Med)**
    - **Median-of-Ratio** (DESeq)
    - **samr** (samr di Bioconductor)
    - **Trimmed Mean of M values (TMM)** (edgeR)
2. **Non-Global Normalization**:
    - **Quantile (Q)**
    - **Reads Per Kilobase per Million mapped reads (RPKM)**
    - **Transcript Per Million (TPM)**

Esempi di normalizzazione

1. **TMM**:
    - Si basa sull'ipotesi che la maggior parte dei geni non siano differenzialmente espressi (DE).
    - Un campione è scelto come riferimento.
    - Calcolo di un fattore TMM come media pesata dei log rapporti tra il campione in considerazione e il riferimento, dopo l'esclusione dei geni più espressi e con log rapporti più grandi.
2. **Median-of-Ratio**:
    - Calcolo del fattore di scala come la mediana del rapporto tra le conte grezze di un campione e la media geometrica tra gli esperimenti del corrispettivo gene.
3. **samr**:
    - Conte grezze divise per il numero totale di reads mappate nel corrispettivo esperimento per un certo insieme di geni non differenzialmente espressi.
    - Moltiplicazione per la media delle conte tra i campioni per quei geni appartenenti all'insieme.
4. **TPM**:

    - Normalizzazione delle conte rispetto alla lunghezza dei trascritti e alla dimensione della libreria.
    - Calcolo delle RPK, somma di tutti i valori RPK in un campione, e divisione per un fattore di scala "per million".

Importanza della normalizzazione

- Facilitare la comparazione delle proporzioni di reads che mappano su un gene tra esperimenti.
- Assicurare che la maggior parte dei geni siano equivalenti espressi nei campioni.
- Risolvere altre distorsioni, come la diversa percentuale di GC contenuta nelle reads.

Scopo dell'analisi

- Identificare i geni con cambiamenti significativi nei livelli di espressione tra le condizioni di studio.

Test statistico

- Determinare se una differenza osservata nelle conte delle reads tra due condizioni è significativa o dovuta a variazione casuale.

Procedura statistica

- Individuare i geni differenzialmente espressi (DE).
- Molti metodi sono in sviluppo; non c'è consenso su quale sia il migliore.

Valutazione comparativa

- Confronto di 8 metodi per l'analisi di differenziale espressione RNA-Seq:
    - DESeq2
    - edgeR
    - baySeq
    - EBSeq
    - ShrinkSeq
    - SAMseq
    - NOISeqBIO
    - voom(+limma)

Obiettivi della valutazione

- Individuare un metodo o un insieme di metodi che funzionano bene in tutte le condizioni sperimentali.

- Fornire linee guida per la scelta del metodo da utilizzare basate su condizioni sperimentali (numerosità campionaria, eterogeneità del campione, ecc.).

Pre-processamento dei dati

- Normalizzazione TMM applicata dove possibile.
    - TMM forniva risultati più soddisfacenti rispetto a tutte le metriche usate in una recente comparazione di metodi di normalizzazione.
    - La median-of-ratio si comporta in modo simile alla TMM.

Risultati della normalizzazione

- Cambiare i metodi di normalizzazione di default con TMM non comporta un significativo impatto sui risultati delle analisi.
- Le differenze rilevate con i vari pacchetti di analisi delle differenziale espressione sono dovute alle caratteristiche degli algoritmi di detection, non ai metodi di normalizzazione.

Conclusioni

- La scelta del metodo di analisi deve considerare le condizioni sperimentali specifiche.
- La normalizzazione TMM è efficace e può essere utilizzata senza influenzare negativamente i risultati.

Metodi per l'analisi della differenziale espressione

Tutti i criteri utilizzano una matrice di conte che contiene il numero di reads che mappano su ciascun trascritto in ciascun campione dell'esperimento.

Di tali metodi sette lavorano direttamente sulle conte (DESeq2, edgeR, baySeq, EBSeq, ShrinkSeq, NOISeqBIO e SAMseq), mentre uno trasforma le conte usando poi il pacchetto R limma, che è stato sviluppato in origine per l'analisi della differenziale espressione per dati di microarray (voom).

I metodi che lavorano direttamente sulle conte possono essere divisi in:

• parametrici: DESeq2, edgeR, baySeq, EBSeq, ShrinkSeq;
• non parametrici: SAMseq, NOISeqBIO

I modelli parametrici, a parte ShrinkSeq, usano un modello Binomiale Negativo (NB) per tenere conto della sovra-dispersione, mentre ShrinkSeq permette all'utente di scegliere tra una varietà di distribuzioni tra cui la Binomiale Negativa e la zero-inflated NB, ovvero una distribuzione che tiene conto del gran numero di conte nulle nei campioni. DESeq2 e edgeR prevedono un classi- co test d'ipotesi mentre gli altri metodi parametrici

utilizzano un approccio Bayesiano. I due metodi non parametrici qui valutati (SAMseq e NOISeq- BIO) non assumono alcuna distribuzione particolare per i dati. Infine, l'approccio di trasformazione voom (dal pacchetto limma di R) ha lo scopo di trovare una trasformazione delle conte per renderle più adatte all'analisi con i metodi tradizionali sviluppati per l'analisi di differenziale espressione per i microarray.

Filtraggio indipendente→**DESeq2** usa la media dell'espressione di ciascun gene tra tutti i campioni come criterio di filtraggio e omette tutti i geni la cui media delle conte normalizzate è sotto una certa soglia derivata dall'aggiustamento per test multipli; di default tale soglia è scelta per massimizzare il numero di geni trovati in base all'FDR specificato dall'utente. Il filtraggio riduce la perdita di potenza dovuta all'aggiustamento per test multipli e non compromette la distribuzione della statistica test poichè, sotto l'ipotesi nulla, questa è marginalmente indipendente dalla statistica di filtraggio

Il false discovery rate (FDR) è un metodo di concettualizzazione del controllo del tasso di errore di I tipo quando sono condotti test multipli. La procedura di controllo dell'FDR, in questo contesto, prevede il controllo della proporzione attesa di falsi positivi tra i geni che sono rilevati come differenzialmente espressi. Tale sistema prevede un controllo dell'errore di I tipo meno restrittivo rispetto alle procedure basate sul family wise error rate (FWER), come la correzione di Bonferroni [54], che controllano la probabilità di avere almeno un falso positivo tra tutti i test fatti. La procedura di controllo dell'FDR ha una potenza superiore, al costo di un aumento del tasso di errori del I tipo.
L'FDR non è una quantità che può essere calcolata, ma va stimata dato che, generalmente, tra i geni rilevati come significativamente DE non si conosce quanti di questi siano realmente DE e quanti no.

In fase di analisi, per determinare la lista dei geni differenzialmente espressi per ogni metodo si sono selezionati quei geni il cui FDR è risultato inferiore a 0.05. Tuttavia, per evitare biases dovuti al metodo di stima dell'FDR adottato dai vari approcci, e per validarne unicamente l'algoritmo di stima della differenziale espressione, si è scelto di considerare anche le liste dei top500.